



دانشکده مهندسی
کامپیوتر و فناوری اطلاعات



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

داده کاوی

(بهار ۱۴۰۱)

تمرین دوم

محمد چوپان ۹۸۳۱۱۲۵

بخش تئوری :

سوال اول :

سوال اول

یک مجموعه داده از حیوانات مختلف به همراه ویژگی‌هایشان را در اختیار داریم، می‌خواهیم با استفاده از روش‌های خوشبندی میزان شباهت هر دو حیوان به هم را از ۱ (کمترین) تا ۳ (بیشترین) مشخص نماییم. برای مثال میزان شباهت شیر و پلنگ $\underline{3}$ و میزان شباهت شیر و گوسفند $\underline{1}$ می‌تواند باشد. الگوریتمی ارائه دهید که این امر را به صورت غیرنظرارت شده^۱ ممکن سازد.

پاسخ :

برای انجام این کار ابتدا می‌توانیم داده‌هارا پیش پردازش کنیم و داده‌ها را به صورت ماتریسی در بیاوریم. سپس برای انتخاب ویژگی می‌توانیم داده‌های غیر عددی را به داده‌های عددی در بیاوریم و ویژگی‌های مهم آن را انتخاب کنیم در صورت نیاز می‌توانیم از الگوریتم کاهش بعد غیر نظارتی PCA نیز استفاده کنیم. در نهایت می‌توانیم با استفاده از الگوریتم خوشبندی kmeans داده‌های خود را دسته بندی کنیم. می‌توانیم K را در اینجا $\underline{3}$ در نظر بگیریم. حال شباهت بین داده‌های یک خوش را $\underline{3}$ در نظر می‌گیریم برای به دست آوردن شباهن بیت داده‌های $\underline{2}$ خوش مختلف نیز می‌توانیم معیار فاصله را در نظر بگیریم و یک عدد بین ۱ تا $\underline{2}$ به آن اختصاص دهیم. یک فرایند دیگر نیز این است که ابتدا ماتریس ویژگی‌ها را تهیه کنیم و پس از مراحل پیش پردازش مانند استاندارد سازی از الگوریتم‌های کاهش بعد مانند PCA استفاده کنیم و در نهایت شباهت بردارهای ویژگی را با استفاده از معیاری مانند فاصله اقلیدسی حساب کنیم و داده‌ها را بین ۱ تا $\underline{3}$ نرمال سازی کنیم.

سوال دوم :

سوال دوم

می‌دانیم که در الگوریتم خوشبندی برای تابع مجاورت^۲ موارد مختلفی را می‌توان استفاده کرد، در موارد زیر اثبات نمایید که نقطه نهایی که به عنوان مرکز انتخاب می‌شود چه نقطه‌ای است. (در رابطه زیر D مجموعه تمامی نقاط داده و C مجموعه تمامی مراکز خوشه‌ها می‌باشد).

$$\sum_{d \in D} \sum_{c \in C} f(d, c)$$

- ($f(d, c) = |d - c|$) • نرم ۱
- ($f(d, c) = \|d - c\|_2^2$) • نرم ۲

ODYSSEY

ODYSSEY ENGINEERING CO., LTD.
514 EISENHOWER AVE., TEHRAN 13
TEL. 963282

$$\sum_{\delta \in D} \Delta c_i = \frac{\partial \text{Cost}}{\partial c_1} \quad \text{بسن تابع}$$

مزینهای مشتبه
باشد

$$= \left[\begin{array}{c} s g(\Delta c_1, -c_1) \\ \vdots \\ \vdots \end{array} \right] \quad \left[\begin{array}{c} \vdots \\ \vdots \end{array} \right]$$

$$(\Delta c_1 - c_1) + (\Delta c_2 - c_2) + \dots = 0$$

یکی در ماست که کوچکتر مسادی به ناگفته است که اعداد زیر متساوی کوچکتر نباشند.

$$\frac{\partial \text{Cost}}{\partial c_1} = \sum_{\delta \in D} \frac{\partial c_1 - \delta}{\partial c_1}$$

کوچکتر از مسادی است

$$\Rightarrow \sum_{\delta \in D} (\Delta c_1 - \delta) = 0 \Rightarrow \Delta \text{Cost}(D, c_1) = \Delta c_1 \in \sum_{\delta \in D}$$

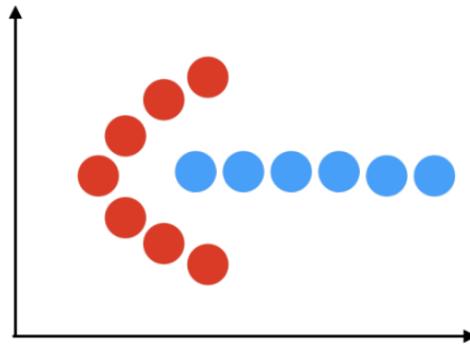
یعنی $c_1 = \min(\Delta)$



سوال سوم:

سوال سوم

الف) فرض کنید داده‌های زیر را می‌خواهیم به ۲ دسته مختلف دسته‌بندی کنیم، پیش‌بینی شما از اجرا الگوریتم k-means را از داده‌های زیر بیان کنید و علت این پیش‌بینی را هم ذکر نمایید.



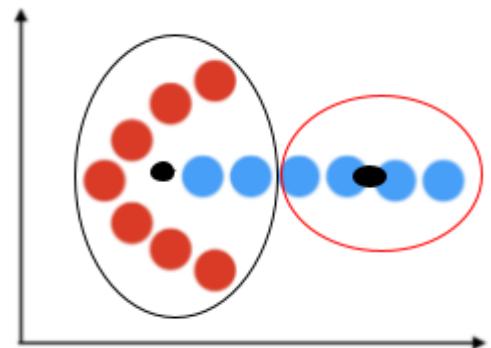
ب) آیا استفاده از روش DBSCAN میتواند برای داده‌های بالا عملکرد بهتری داشته باشد؟ علت را توضیح دهید.

ج) توضیح دهید در چه زمانی خوشبندی بر مبنای چگالی عملکرد مناسبی نخواهد داشت؟ مثال بزنید.

پاسخ :

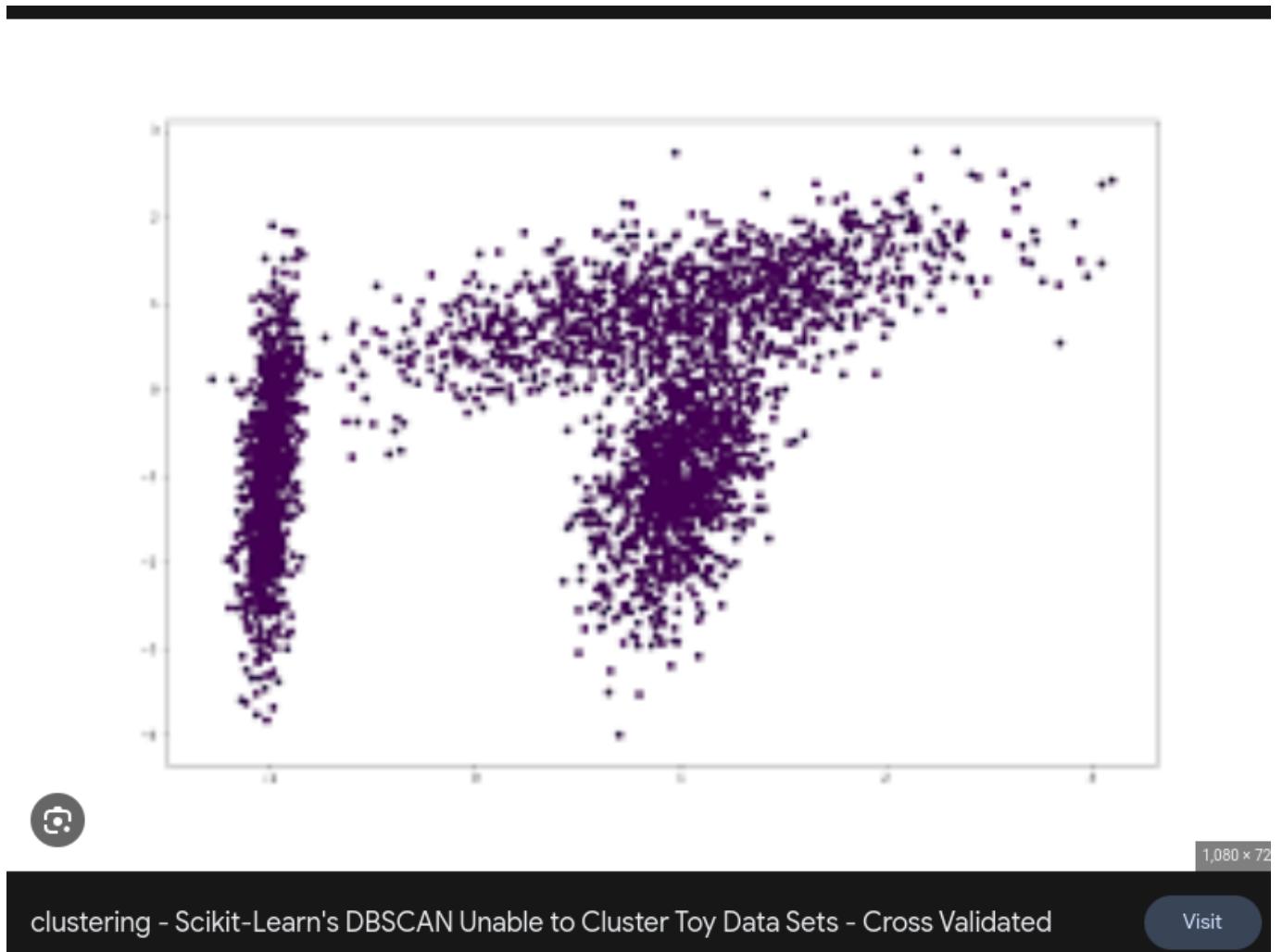
پیش‌بینی ما مانند شکل زیر است به دلیل اینکه سعی می‌کند بیشترین نقاط در کمترین فاصله را داشته باشد که عملکرد خوب و درستی هم نیست. پیاده سازی انجام شده نیز پیش‌بینی ما را تایید می‌کند. همچنین اگر مرکز ابتدایی هر رندومی باشند به وسط دو دسته میل پیدا می‌کند و حرف ما را تایید می‌کند.

که دو نقطه کوچک شکل مرکز نهایی ما هستند.



ب: بله زیرا روش DBSCAN یک روش مبتنی بر چگالی می‌باشد و اگر EPS را مقدار معقولی قرار دهیم یعنی بیشتر از حداقل فاصله آبی‌ها با هم و قرمز‌ها با هم و کمتر از حداقل فاصله آبی و قرمز با یکدیگر می‌تواند این دو دسته را به خوبی جدا کند. این الگوریتم به دلیل اینکه با فاصله مرکز کار نمی‌کند و از یک مقدار heuristic استفاده می‌کند در شکل‌هایی که این حالت را دارند مانند دو دایره تو در تو عملکرد بهتری دارد.

زمانی که چگالی داده های ما در نقاط متفاوتی درون یک خوشه متفاوت باشد یعنی درون یک خوشه در یکجا خیلی زیاد و در یکجا خیلی کم باشد و به اصلاح توضیع متوازنی نداشته باشد. مانند شکل زیر یا زمانی که داده های ما خیلی بعد های متفاوتی داشته باشد نمیتواند به خوبی عمل کند.

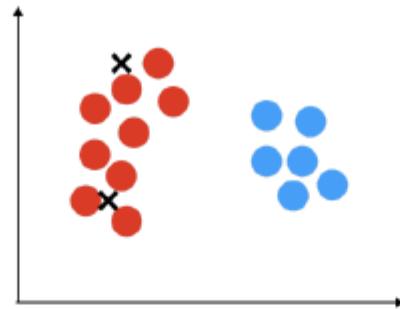


به دلیل مقدار زیاد noise و outlier علی رغم اینکه DBSCAN در نویز ها خوب عمل میکند اما اینجا موفق نشده است.
در کل الگوریتم های مبتنی بر چگالی این مشکل را دارند.

سوال چهارم :

سوال چهارم

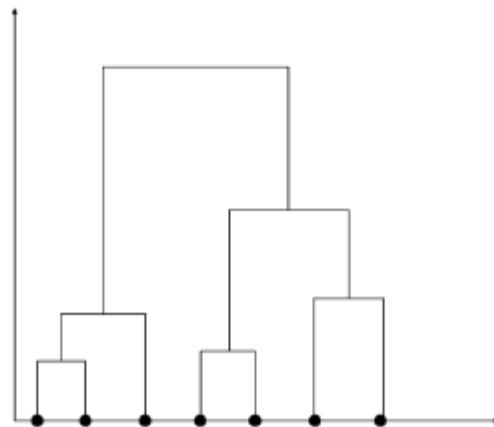
الف) نتیجه اعمال الگوریتم k-means را بر روی داده‌های زیر مشخص کنید. (ضرب در بیانگر مراکز اولیه است)



ب) برای حل مشکل بالا از راهکارهای گوناگونی استفاده می‌شود در رابطه با هر یک از این راهکارها را تحقیق کرده و مزایا و معایب آن‌ها را توضیح دهید

- استفاده از medoid به جای median
- انتخاب نقاط اولیه به شکلی که بیشترین فاصله را از هم داشته باشند
- انتخاب نقاط اولیه بر اساس توزیع داده‌ها
- انتخاب چندباره مراکز اولیه برای رسیدن به جواب مناسب

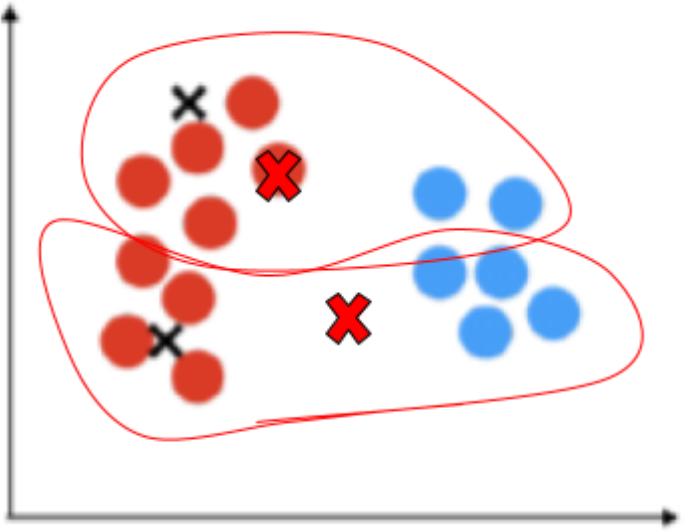
ج) دندروگرام زیر، انجام خوشبندی سلسله مراتبی را بر روی یک مجموعه دادگان را نشان می‌دهد، با توجه به دندروگرام مشخص نمایید که اگر بخواهیم بر روی داده‌های زیر الگوریتم k-means را اجرا نماییم بهتر است که چه مقداری را به k دهیم.



پاسخ :

الف:

نتایج برای الگوریتم $k=2$ برابر است با شکل زیر دلیل آن هم این است که مرکز پایینی وسط ۴ تا آبی و ۴ تا قرمز است و مرکز بالایی نیز به دلیل بیشتر بودن قرمز‌ها به سمت قرمز‌ها مایل شده است. علت این اتفاق نیز نزدیک بود مرکز اولی و درست انتخاب نشدن آن‌ها است.



: ب

استفاده از medoid به جای median :

مزایا : انتخاب مرکز از medoid ها از این نظر بهتر است که داده های خارجی یا همان outlier ها را بهتر تشخیص داده و نسبت به آن ها کمتر حساس است. و اینکه مدoid به دلیل اینکه از خود داده ها است برای داده های غیر عددی نیز مناسب تر است. همچنین استفاده از مدoid باعث پایداری الگوریتم می شود و تاقیرات نویز ها را نیز کمتر می کند.

معایب : محاسبات آن بیشتر است و از نظر محاسبات هزینه زیاد بر ما تحمیل می شود و بعضی وقتی تشخیص خود مدoid دنیز سخت است.

انتخاب نقطه اولیه به شکلی که بیشترین فاصله را داشته باشند:

مزایا : برای داده هایی که خیلی با هم فاصله دارند مناسب تر است و همچنین در بهینه های محلی گیر نمیکند لزوماً

معایب : به داده های بیرونی و نویز ها بسیار حساس تر می شود و همچنین باعث میشود که تعداد تکراری بیشتری داشته باشیم. و اگر خوشه های ما با هم مقداری همپوشانی داشته باشند نتایج جالبی نخواهد داشت.

انتخاب نقاط اولیه بر اساس توزیع داده ها :

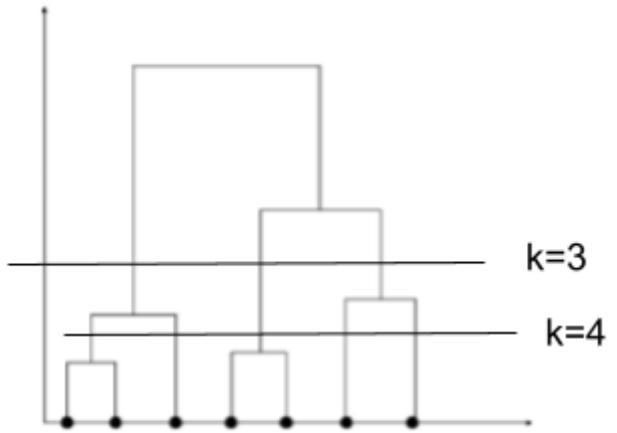
مزایا : احتمال اینکه در مینیمم محلی گیر کنیم کمتر است و معمولاً برای داده های است که یک توزیع یکسانی ندارند

معایب : محاسبات بسیار زیادی دارد و مجدد بر نویز ها نیز حساس تر می شود و همچنین این حالت وجود دارد که به علت اینکه داده ها توزیع یکسانی ندارند در خوشه های خاصی مرکز بیشتری داشته باشد.

انتخاب چند باره مراکز اولیه برای رسیدن به جواب مناسب :

مزایا : یکی از بهترین های روش ها این است به دلیل اینکه هم خط را کمتر میکند هم نقش نقاط بیرونی را کمزنگ تر میکند و اینکه تقریباً احتمال گیر کردن در مینیمم محلی وجود ندارد

معایب : تعداد محاسبات ما خیلی بالا می رود و یک روش مشخص برای اتمام کار وجود ندارد لزوماً



بهتر است که یکی از این دو عدد را انتخاب کنیم هم بر اساس روش elbow هم اینکه ارتفاع نمودار دندوگرام میزان شباهت داده ها را نشان میدهد و اینجا هر چی بیشتر باشد داده ها متفاوت ترند برای دسته بندی این اعداد نتایج بهتری میدهند. بین ۳ یا ۴ یک داده تنها تفاوت دارد که میتوان ۳ را برای محاسبات کمتر نیز انتخاب کرد.

سوال پنجم :

سوال پنجم

ماتریس زیر را در نظر بگیرید با استفاده از روش PCA داده ها را به یک بعد انتقال داده و ماتریس داده حاصل را بدست آورید.

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 2 & 1 \\ -1 & -1 \\ -1 & -2 \\ -2 & -1 \end{bmatrix}$$

پاسخ :

ابتدا مرحله اول باید داده ها را استاندارد کنیم به دلیل اینکه میانگین ما برابر ۰ است نیازی به انجام این مرحله نداریم. در مرحله بعدی باید ماتریکس کواریانس را محاسبه کنیم که بقیه محاسبات در تصویر زیر وجود دارد. در پاسخ نهایی تفاوتی ندارد اما مخرج cov ابتدا n بود سپس به $n-1$ تغییر دادم.

ODYSSEY

ODYSSEY ENGINEERING CO., LTD.
514 EISENHOWER AVE., TEHRAN 13
TEL. 963282

$$\text{Cov}_{n,g} = \frac{\sum (x_i - \bar{x})(g_i - \bar{g})}{N-1}$$

$$\text{Var}(x) = \frac{\sum (x_i - \bar{x})^2}{N}$$

$$\text{Var}(g) = \sigma^2$$

$$\text{Cov}(x, g) = \frac{1}{N} \cdot \sigma^2$$

$$\text{Cov}_{n,n} = \begin{bmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{bmatrix}$$

$$|\lambda I - A| = \begin{vmatrix} \lambda & 1 \\ 1 & \lambda \end{vmatrix} - \begin{bmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{bmatrix} = \begin{vmatrix} \lambda - \sigma^2 & \sigma^2 \\ \sigma^2 & \lambda - \sigma^2 \end{vmatrix}$$

$$\Rightarrow (\lambda - \sigma^2)^2 - \sigma^4 \Rightarrow (\lambda - \sigma^2)^2 = 0 \Rightarrow \lambda = \pm \sigma^2$$

$$\begin{bmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \sigma^2 \begin{bmatrix} u \\ v \end{bmatrix} \Rightarrow \frac{1}{\sigma^2} u + v = 1 \Rightarrow u = \frac{1}{\sigma^2} - v$$

$$u \quad u \quad \sigma^2 u \Rightarrow \frac{1}{\sigma^2} u + v = 1 \Rightarrow v = \frac{1}{\sigma^2} - u$$

$$\frac{\sigma^2 + \sigma^2}{\sigma^2 + \sigma^2} = 1 \quad \text{مساواه میان فاصله های متریک}$$

$$\text{final DS} = F \sqrt{a \cdot \text{rising}} = [1 \ 1]$$

$$\Rightarrow (1 \ 1 \ 1 -1 -1 -1)$$



سوال ششم:

سوال ششم

با فرض آستانه پشتیبانی τ برابر 0.3 و آستانه اطمینان δ برابر 0.4 ، مجموعه تمام قوانین انجمنی ممکن را بنویسید

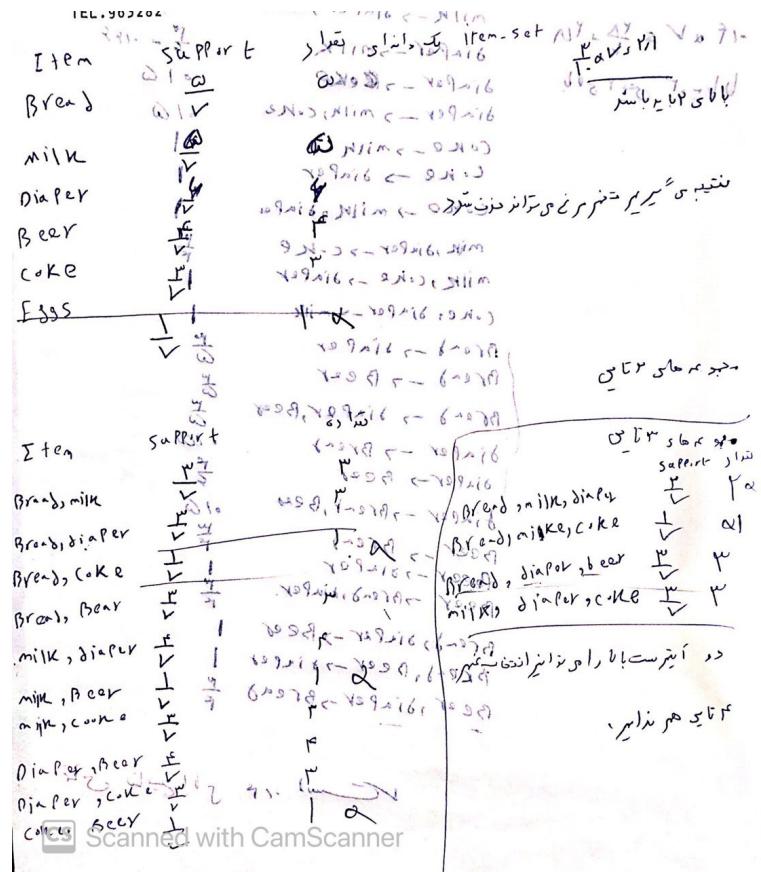
TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke
6	Milk, Diaper, Coke
7	Bread, Diaper, Beer

پاسخ :

برای اینکه support بالای $3/0$ را حساب کنیم میتوانیم از این قانون استفاده کنیم :

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

که ابتدا از مجموعه های یک عضوی شروع کنیم و در نهایت آن هایی که بالای $3/0$ هستند را حساب کرده و آستانه اطمینان بالا $4/0$ را انتخاب کنیم:



ادامه پاسخ در عکس

Bread, diaper, Beer	Rule	Confidence
milk, diaper, coke	milk \rightarrow diaper	0.99
	milk \rightarrow Beer	0.99
	milk \rightarrow diaper, Beer	0.99
	diaper \rightarrow milk	0.99
	diaper \rightarrow Beer	0.99
	diaper \rightarrow milk, coke	0.99
	coke \rightarrow milk	0.99
	coke \rightarrow diaper	0.99
	coke \rightarrow milk, diaper	0.99
	milk, diaper \rightarrow coke	0.99
	milk, coke \rightarrow diaper	0.99
	coke, diaper \rightarrow milk	0.99
	Bread \rightarrow diaper	0.99
	Bread \rightarrow Beer	0.99
	Bread \rightarrow diaper, Beer	0.99
	diaper \rightarrow Bread	0.99
	diaper \rightarrow Beer	0.99
	diaper \rightarrow Bread, Beer	0.99
	Bread \rightarrow Beer	0.99
	Bread \rightarrow diaper	0.99
	Bread, Beer \rightarrow diaper	0.99
	Bread, diaper \rightarrow Beer	0.99

کل نتایج با از ۱۳۰۰

Scanned with CamScanner

سوال هفتم :

سوال هفتم

با فرض آستانه پشتیبانی $\frac{1}{3}$ و آستانه اطمینان $\frac{2}{3}$ ، مجموعه آیتم‌های پر تکرار را به دست آورید. در مرحله بعد مجموعه‌ی تمام قوانین انجمانی ممکن را به دست آورید.

ترکیش	آیتم
T1	{a, b, c}
T2	{d, c}
T3	{a, b, c}
T4	{d, b}
T5	{a, e}
T6	{a, e, d}
T7	{b, c}
T8	{a, b, c, d}

